BAYESIAN MODEL AVERAGING

Sibghat Ullah & Siddhant Tandon

SAPIENZA UNIVERSITY OF ROME MS Data Science

Table of Contents

Abstract

Standard Statistical practices, too often than not, ignore the model uncertainty. A statistical model is typically built to explain the behavior of observable data. Which means that it is rarely perfect. And because of that reason, model uncertainty is one of the most interesting areas for statisticians to explore. Typically, a Data Scientist picks a model from a famous class of models and proceeds as if it is the actual model which generated the data. Proceeding with this approach leads to many problems specifically leading to over confident inferences and predictions. In this report, we would like to suggest a solution to this strategy. A concept called "Bayesian Model Averaging". As it is self-explanatory, it can only be implemented with a Bayesian approach. There are other nice techniques which could help us solve the problem of Model Uncertainty. However, in this study we would be focusing only on "Bayesian Model Averaging" aka "BMA".

Introduction of BMA

We can start by taking an example of a survival analysis case. Suppose our potential statistician has chosen proportional hazard regression model to analyze a data of cancer patients where he has gathered the demographics of patients and medical covariates and aims to predict the survival time for future patients. Suppose he chooses the medical covariates that could be important for this analysis and comes with a statistical model *M**. He verifies that this model fits the data in an effective way and parameter estimation is sensible. After this, he uses this model for predictions, inferences and confidence intervals.

The problem with this strategy is, suppose there's another model M** that also fits the data in an effective way but leads to relatively different inferences and predictions. Previous studies on model uncertainties have indicated that in this case, choosing the model M* alone and making future inferences and predictions using that model is very risky and could backfire.

Thus, we should use a nice technique called "Bayesian Model Averaging". Suppose Δ is the quantity of interest such as a future observable or utility of a course of action, then the posterior distribution for this quantity given the data "D" is:

$$pr\ (\Delta \mid D) = \sum_{k=1}^{K} pr(\Delta \mid Mk, D)\ pr\ (Mk \mid D) \qquad (1)$$

Which is an average of the posterior distributions for each of the models considered, and weighted by their posterior model probability. The posterior probability for a model M_k is given as:

$$pr\ (Mk \mid D) = \frac{pr(D \mid Mk)\ pr(Mk)}{\sum_{l=1}^{K} pr(D \mid Ml) pr(Ml)} \qquad (2)$$

Where

$$pr(D \mid Mk) = \int pr(\ D \mid \theta k, \mathrm{Mk}) pr\ (\theta k \mid Mk) \mathrm{d}\theta k \qquad (3)$$

Is the integrated likelihood of model Mk. θk is vector of parameters for model Mk and pr(Mk) is the prior probability that model Mk is the true model. These probabilities are implicitly conditional on ***M***, which is the set of all models being considered.

Then posterior mean and variance of Δ using BMA are as under:

$$E[\Delta \mid D] = \sum_{k=0}^{K} \breve{\Delta}\mathrm{k}\ \mathrm{pr}\ (\mathrm{Mk} \mid \mathrm{D})$$

And

$$Var\ [\Delta \mid D] = \sum_{k=0}^{K} (Var[\Delta \mid D, Mk + \breve{\Delta}\mathrm{k}^2])\ pr\ (Mk \mid D) - E[\Delta \mid D]^2$$

Where

$$\breve{\Delta}\mathrm{k} = \mathrm{E}[\Delta \mid D, Mk]$$

It has been theoretically proved that averaging over all the models like this results in more accurate predictions and better inference. Furthermore, it has also been proven empirically. As simple as BMA might sound, there're certain issues associated with its implementation. A brief description of each such issue is hereunder.

- The number of terms in (1) is generally enormous, thus making the summation impractical.
- The implicit integrals in (1) are always hard to compute. MCMC can help but it has a lot of technical issues as well.
- Specification of prior probability for model Mk, over all the other competing models is a serious challenge.
- Choosing the class of model, is also a challenging task in general.

BMA Implementation

The above-mentioned issues and their solutions will be discussed in detail.

Managing the Summation

The number of terms in the first equation are too high, generally and need to be taken care of. There're two approaches to solve this issue.

The first approach is to average over a subset of models which're supported by the data. The Occam's windows method is a wonderful way to select such models. The Occam's window method follows two principles.

First is that if a model performs very poor compared to a benchmark model, it needs to be dropped. Thus, the models belonging to the equation

$$A = \left\{ Mk\colon \frac{\max l\ \{pr\ (Ml \mid D)\}}{pr\ (Mk \mid D)} \leq \mathrm{C} \right\} \qquad (4)$$

Should be excluded from (1) and C can be chosen by the statistician.

Second principle, which is an optional one, is to exclude complex models which receive less support from data than their simpler version. So, this leads to exclude the models belonging to

$$\mathrm{B} = \left\{ \mathrm{Mk}: \exists\, Ml \in \mathrm{A}\,, Ml \subset \mathrm{Mk}, \frac{pr\,(Ml\,|D)}{pr\,(Mk\,|D)} > 1 \right\} \qquad (5)$$

And (1) is thus replaced by

$$pr\,(\Delta\,|D) = \sum_{Mk \in Z} pr(\Delta\,|Mk, D)pr(Mk|D) \qquad (6)$$

Where

$$Z = A \setminus B$$

And all probabilities are implicitly conditional on set of models in Z.

Now, we need to identify which models to be put in Z. For that, we'd use an algorithm which is described below.

Occam's Algorithm to select Models

Let A=$\emptyset$ and C=set of all models considered, Then

- Select a model M from C
- C <- C \ {M} and A <- A U {M}
- Select a sub model M0 of M. In graphical models, this can be thought of as removing link(s) from the model M.
- Compute

$$B = log\{\frac{pr(M0\,|D)}{pr\,(M\,|D)}\}$$

- If B > O_R, then A= A \ {M} and if M0 $\notin$ C, then C <- C U {M0}
- If $O_L \leq B \leq O_R$, then if M0 $\notin$ C , C <- C U {M0}
- If there're more sub models of M0, then we must go to step 3.
- If C$\neq\emptyset$, then go to step 1.

Interpretation of Occam's algorithm

The above model selection algorithm can be interpreted by these three statements.

1. If the log posterior odds (B in the above algorithm) is positive, then we reject model M and rather consider M0. This could be achieved by requiring the log posterior odds to be greater than a positive constant O_R
2. If the log posterior odds are negative but small, providing evidence against smaller model M0, we consider both models and neither one is rejected.
3. If the log posterior odds are large and negative, meaning O_L = -log (C) where C is chosen from (5), then we reject M0 and consider M.

This sums up the first approach to be used for summation. Below would be discussed the second approach for summation.

The second approach focuses on Markov chains. More specifically, it can be said that Markov Chain Monte Carlo Model Composition aka MC^3 uses Markov Chain to directly approximate (1). Let ***M*** denote the space of models under consideration. We can come up with a Markov Chain {M(t)}, t=1,2…. With state space ***M*** and stationary distribution pr(Mi | D) and simulate the Markov Chain to obtain observations M(1)……M(N). Then for any function f(Mi) defined on ***M,*** the average

$$\check{F} = \sum_{t=1}^{N} f(M(t)) \frac{1}{N}$$

Is an estimate of E(f(M)). It can be proved theoretically that this converges as N goes to infinity. To construct a Markov Chain like this, we should define a neighborhood of M for each M ∈ ***M***. A Typical example of such neighborhood in graphical models would be a model with one more link or a model with one less link, plus the model M itself as well. We can define a transition matrix q by setting

$$q(M \to M') = 0 \; for \; all \; M' \notin \mathrm{nbd(M)}$$

And

$$q(M \to M') \neq 0 \; for \; all \; M' \in nbd(M)$$

When the Markov Chain is in state M, we can step next by drawing M' from q. M' is accepted with probability

$$min \left\{1, \frac{pr(M'|D)}{pr\ (M\ |D)}\right\}$$

We can use Metropolis Hastings Algorithm to construct such a Markov Chain.

Studies have shown that this second approach has done decently well, and for graphical models, it can be computationally sped up. However, please note that the typical issues of MCMC always remain. These typical issues are Convergence, Diagnostics, Autocorrelation and computation time. Furthermore, technical expertise on MCMC can't be achieved by every statistician and requires vast experience and knowledge. Having said that, both approaches are good. One doesn't have a lot of evidence over the other, and as such it depends upon the Statistician to use whatever approach he thinks would perform better in a certain case.

Computing Integrals in BMA

Computing the integrals is always a challenging task. In most of the situations, approximating the Integrals is often the solution. Studies done on graphical models and linear regression have shown that close form integrals, for the marginal likelihood in (3), are available and thus can be used. Most of the work however, done to study BMA and integrals computation results in approximation. We'll be discussing these approximation in detail in the next section.

Computing Integrals for Linear Regression

The selection of independent variables in Linear Regression is an important task. The objective of the independent variables is that we should find the best model of the form

$$Y = B0 + \sum_{j=1}^{p} Bij\ Xij + \varepsilon$$

Where Xi1…., Xip is a subset of X1……Xk.

Bayesian Model Averaging however tries to average over all possible set of predictors. Our conclusive results on this issues after rigorous studies also include Transformations and outliers. So basically, we would use Box-Cox class of power transformations for the response. This would result in estimating the parameters. The statistical model is

$$Y(\mathrm{p}) = XB + \ \varepsilon \text{ where } \varepsilon \sim \mathrm{N}(0, \sigma 2\mathrm{I})$$

The transformation for the response would be as

$$Y(p) = \begin{cases} \dfrac{y^p - 1}{p} & p \neq 0 \\ \log(y), & p = 0 \end{cases}$$

To have the meaningful power transformation so that it has biological or physical interpretation, we can assume the values of p to be (-1,0,0.5,1) and average over these values.

For the transformation of independent variables, we can use Alternate Conditional Expectation Algorithm and then we can use change point transformations. For the sake of simplicity, the ACE algorithm is not discussed here. It can be seen in many research articles and can be used to select the most optimal transformations of independent variables which can then be used to approximate the integral in this case.

Computing Integrals for Generalized Linear Models

A generalized linear model consists of independent variables, a link function and a variance function. Thus, each possible combination of these factors results in a different generalized linear model. Previous studies have shown on how to compute the Bayes factor for generalized linear models. The Bayes factor for a model M1 against another model M0 is given as:

$$B10 = \frac{pr\ (D\ |M1)}{pr\ (D\ |M0)}$$

In other words, it's the ratio of the marginal likelihoods. The Bayes factor, in turn, yields posterior model probabilities for all the models, and enable BMA.

Suppose that k+1 models (M0, M1, ……, Mk) are being considered. Each of model M1, …., Mk is compared with M0 yielding Bayes factor B10, ….m Bk0. Then the posterior probability of model Mk is given as

$$pr(Mk\ |D) = \frac{\alpha k * Bk0}{\sum_{r=0}^{K} \alpha r * Br0} \qquad (7)$$

Where

$$\alpha k = \frac{pr(Mk)}{pr(M0)}$$

Now suppose that Yi is a dependent variable and that Xi = (Xi1…. Xip) is the vector of independent variables for i = 1…. n. A generalized linear model is defined by specifying pr(Yi | Xi, B) in such a way that E[Yi | Xi] = μi, Var[Yi|Xi] = σ2var(μi) and g(μi) = Xi B, where B=(B1….Bp)^T , and here g is known as the link function. Here X is the n*p matrix of covariates where Xi1=1, and it is assumed that σ2 is known.

Considering the Bayes factor for model M0, defined by setting Bj = 0 (j=2…. p) against M1. The likelihoods for M0 and M1 can be explicitly mentioned and after specifying the prior, the Laplace approximation and then the Newton Raphson method yields.

$$2logB10 \approx \chi 2 + (E1 - E0) \qquad (8)$$

Furthermore,

$$\chi 2 = 2\{L1(\widetilde{B1}) - L0(\widetilde{B0})\}$$

And

$$Lk(\widetilde{Bk}) = \log(pr(D\ |Bk, Mk))$$

Is the log likelihood when M0 is nested inside M1 and $\chi 2$ is the standard likelihood ratio test statistics. Also,

$$Ek = 2\lambda k(\widetilde{Bk}) + \lambda' k * transpose(\widetilde{Bk}) * inverse(Fk + Gk)\{2 - Fk * inverse(Fk + Gk)\} * \lambda' k(\widetilde{Bk}) - \log|Fk + Gk| + pk\log(2\pi)$$

Where Fk is the expected fisher information matrix, And Gk=inverse (Var[Bk | Mk]) and $\lambda k(\widetilde{Bk}) = \log pr(Bk\ |Mk)$ is the log prior density and $\lambda' k(\widetilde{Bk})$ is the pk derivates of λk(Bk) with respect to the elements of Bk (k=0,1)

The relative error regarding this approximation is

$$O\left(\frac{1}{\sqrt{n}}\right)$$

However, can be reduced by using canonical link function.

Survival Analysis

In Survival analysis, the most important thing is modeling the hazard rate. Typically, Cox proportional hazard model is used for this purpose. The Cox model specifies the hazard rate for subject I with covariate vector Xi to be

$$\lambda(\mathrm{t}\,|\mathrm{Xi}) = \lambda\mathrm{o}\,(\mathrm{t}) * \exp(Xi\ B) \qquad (9)$$

where λo (t) is the baseline hazard function at time t, and B is a vector of unknown parameters. The estimation of B is commonly based on partial likelihood.

$$Pl(B) = \prod_{i=1}^{n} \left(\frac{\exp(XiB)}{\sum_{l \epsilon Ri} \exp(transpose(Xl) * B)} \right)^{wi}$$

Where Ri is the risk set at time t and wi is the indicator whether patient i is censored.

The result about survival analysis are based on approximations. The following MLE approximation is used.

$$pr(\Delta\,|Mk, D) \approx pr(\Delta\,|Mk, \widetilde{Bk}, D)$$

And the Laplace approximation

$$logpr(D|Mk) \approx logpr\left(D \mid \widetilde{Bk}, Mk\right) - dk \log(n) \qquad (10)$$

Where dk is the dimension of Bk where n is usually taken to be the number of cases. Few studies have taken n to be the number of uncensored patients.

To identify the correct models, the Occam's window algorithm is used. Some studies have also used leaps and bound Algorithm which we would skip for the time being. Rest assured, the above equations give us an approximation for the Survival analysis to implement Bayesian Model Averaging.

Graphical Models: Missing Data & Auxiliary variables

A graphical statistical model is a graph that shows the conditional independence among relationships of nodes. These nodes can be thought of as Random Variables. In this study, we'd focus only on acyclic directed graphs.

The example of a typical acyclic directed graph is shown as:

A…….>B…….>C

The joint density of these three variables is given as

$$pr(A, B, C) = pr(A) * pr(B\,|A) * pr(C \mid B)$$

The implementation of BMA in case of graphical models with missing data and auxiliary variables is done through the reexpression of the Bayes factor for two models M0 and M1.

$$\frac{pr(D|M0)}{pr(D\,|M1)} = \boldsymbol{E}\left(\frac{pr(D, Z\,|M0)}{pr(D, Z\,|M1)} \middle| D, M1 \right)$$

Z here shows the missing data and/or auxiliary variables. The above expectation can be approximated by simulating the missing data from its predictive distribution.

Specifying Prior Model Probabilities

One of the most important challenges in Bayesian Model Averaging is specifying the prior distribution for the models. This is important because this could change the inference and results and thus needs to be carefully selected. When there's very little information available regarding the domain of the problem being studied, then all models are important and need to have a uniform prior distribution. However, if the Statistician has expertise over a specific domain then it's perfectly fine to give more probability to a model which is more likely to explain the hidden phenomenon of the observable quantity. Studies have shown that informative prior in Bayesian Model Averaging result in better predictive performance.

Typically, in the case of linear regression or Cox proportional hazard rate, prior probability on model Mi can be specified as:

$$pr(Mi) = \prod_{j=1}^{p} \pi j^{\delta ij} \, (1 - \pi j)^{1-\delta ij} \qquad (11)$$

Where $\pi j \in [0,1]$ is the prior probability that $Bj \neq 0$ in a regression model and δij is an indicator whether variable j is included in the model Mi.

If we assign $\pi j = 0.5$ for all j, then this would mean a uniform prior across model space. If $\pi j < 0.5$, this imposes penalty for large models while if $\pi j = 1$ makes sure that variable j is included in all the models.

As for the graphical models are concerned, it is better to use probability for each link and then multiplying these link probabilities to get the required prior distribution. This prior has been used in many articles and is quite intuitive.

Another approach for eliciting the prior distribution especially in the case of discrete data is that you start with a uniform prior and then receive an imaginary data from the domain expert and update your prior with respect to that data. That, updated prior is then used in Bayesian Model Averaging.

Predictive Performance

One of the fundamental purposes for a statistical model is to predict the observable quantity of interest which would be known in future. Thus, prediction is an important criterion to judge a model. In BMA, we can and should do the same.

To do this, we randomly split the data in two halves. Then we apply each model selection method to the first half of the data which is called the build data. Performance is measured on second half of the data. One of the possible methodology to measure the predictive performance is logarithmic scoring rule of good. This uses, the predictive performance of a model M, using the sum of logarithms of observed ordinates of the predictive density for every observation in test data set.

This is given by:

$$- \sum_{d \in test\_data} log \; pr(d \mid M, training\_data) \qquad (12)$$

And so, in Bayesian Model Averaging this becomes

$$-\sum_{d \in test\_data} log(\sum_{M \in C} pr(d \mid M, training\_data) pr(M \mid training\_data)) \qquad (13)$$

The small the result of a given model average (13) or a model (12), the better predictive performance it would have.

Several other measures exist to explain the behavior of predictive performance for Bayesian Model Averaging but we'd skip these since the above equation provide us a benchmark methodology to measure it.

Example and Software

An example in R, using the packages "BMA" and "iterativeBMAsurv" is also provided by us. Similar packages exist in R and other languages. The example is specific to the Survival Analysis using Bayesian Model Averaging.

Discussion

Bayesian Model Averaging is a nice technique which lets you perform better in terms of inference and predictive performance but the philosophical question here is whether it's worth it or not? As we've seen the difficulties in implementing this strategy, we can have a very strong opinion about it. BMA like every other statistical technique is just a technique, and no technique is better than the other. It just depends upon your own judgement if it can be useful for you or not. Too often, the constraints like time and technical expertise are a huge issue to implement it. Furthermore, the application or the domain doesn't require this. But sometimes, the domain or the specific application require you to be as accurate as possible, and in that case, perhaps BMA is a better solution. Nonetheless, as a statistician having a general idea about this technique is a pleasant thing to have since this can help us a lot in Survival Analysis or Biostatistics in general where accuracy matters a lot.

Acknowledgments

We'd like to thank these authors and their work previously done on this subject.

Bayesian Model Averaging. A tutorial (Jennifer A. Hoeting, David Medigan, Adrian E. Raftey & Chris. Volinsky)

Estimating Optimal Transformations for Multiple egression & Correlation (Leo Brieman & Jerome H. Friedman)

The Iterative Bayesian Model Averaging for Survival Analysis: An Improve method for Gene Selection & Survival Analysis on Microarray data (Amalia Annest, Roger E. Bumgarner, Adrian E. Raftey & Ka Yee Yeung)

Model Selection & Accounting for model Uncertainty in Graphical models using Occam's window (David Medigan & Adrian E. Raftey)